

520.42961X00

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant(s): SUTOH, et al.  
Serial No.: Not assigned  
Filed: July 31, 2003  
Title: DATA CONTROL METHOD FOR DUPLICATING DATA  
BETWEEN COMPUTER SYSTEMS  
Group: Not assigned

LETTER CLAIMING RIGHT OF PRIORITY

Mail Stop Patent Application  
Commissioner for Patents  
P.O. Box 1450  
Alexandria, VA 22313-1450

July 31, 2003

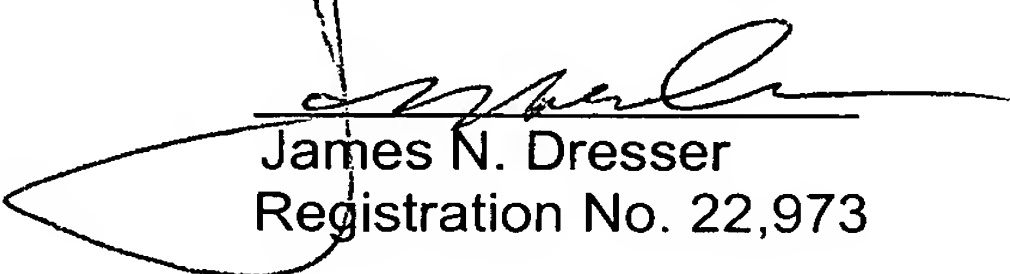
Sir:

Under the provisions of 35 USC 119 and 37 CFR 1.55, the applicant(s) hereby claim(s) the right of priority based on Japanese Application No.(s) 2003-086920 filed March 27, 2003.

A certified copy of said Japanese Application is attached.

Respectfully submitted,

ANTONELLI, TERRY, STOUT & KRAUS, LLP

  
James N. Dresser  
Registration No. 22,973

JND/amr  
Attachment  
(703) 312-6600

日 本 国 ・ 特 許 庁  
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出 願 年 月 日  
Date of Application:

2003年 3月27日

出 願 番 号  
Application Number:

特願2003-086920

[ ST.10/C ]:

[ JP2003-086920 ]

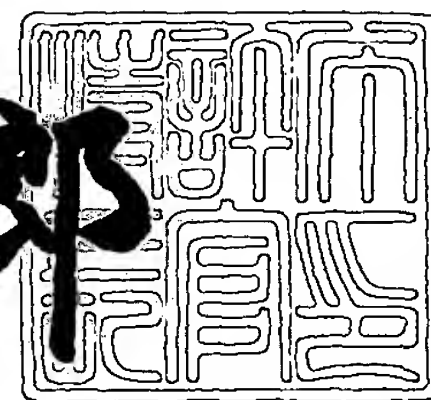
出 願 人  
Applicant(s):

株式会社日立製作所

2003年 6月19日

特 許 庁 長 官  
Commissioner,  
Japan Patent Office

太田信一郎



出証番号 出証特2003-3048011

【書類名】 特許願

【整理番号】 H03001221A

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 12/00

【発明者】

    【住所又は居所】 東京都国分寺市東恋ヶ窪一丁目 2 8 0 番地 株式会社日立製作所中央研究所内

    【氏名】 須藤 敦之

【発明者】

    【住所又は居所】 東京都国分寺市東恋ヶ窪一丁目 2 8 0 番地 株式会社日立製作所中央研究所内

    【氏名】 馬場 恒彦

【特許出願人】

    【識別番号】 000005108

    【氏名又は名称】 株式会社 日立製作所

【代理人】

    【識別番号】 100075096

    【弁理士】

    【氏名又は名称】 作田 康夫

    【電話番号】 03-3212-1111

【手数料の表示】

    【予納台帳番号】 013088

    【納付金額】 21,000円

【提出物件の目録】

    【物件名】 明細書 1

    【物件名】 図面 1

    【物件名】 要約書 1

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 計算機システム間のデータ二重化制御方法

【特許請求の範囲】

【請求項 1】

データを保持する記憶媒体を有する複数の記憶装置と、該記憶装置を制御する制御装置とを有し、前記制御装置に接続された外部装置に前記記憶装置の内の特定の記憶装置のデータが変更されたことを通知する手段を持つことを特徴とするストレージシステム。

【請求項 2】

前記複数の記憶装置の内のデータが変更されたことを前記外部装置に通知すべき記憶装置を外部から選択するインタフェースをさらに有する請求項 1 記載のストレージシステム。

【請求項 3】

データを保持する記憶媒体を有する複数の記憶装置と、該記憶装置を制御する制御装置とを有し、前記制御装置に接続された外部装置に前記制御装置の内の特定の記憶装置の状態が変化したことを通知する手段を持つことを特徴とするストレージシステム。

【請求項 4】

前記複数の記憶装置の内の状態が変化したことを前記外部装置に通知すべき制御装置を外部から選択するインタフェースをさらに有することを特徴とする請求項 3 記載のストレージシステム。

【請求項 5】

ストレージシステムと接続して該ストレージシステムと制御信号およびデータを送受信する接続装置を有し、かつ、該接続装置を介して接続された前記ストレージシステム内の特定の記憶装置のデータが変更されたこと示す通知を受信するインタフェースを持つことを特徴とする計算機システム。

【請求項 6】

前記ストレージシステムに対し、データが変更されたことを示す通知を受信すべき記憶装置を選択して指示するインタフェースを持つことを特徴とする請求項

5 記載の計算機システム。

【請求項 7】

ストレージシステムと接続して該ストレージシステムと制御信号およびデータを送受信する接続装置を有し、かつ、該接続装置を介して接続された前記ストレージシステムから該ストレージシステム内の制御装置の状態が変化したことを受信するインタフェースを持つことを特徴とする計算機システム。

【請求項 8】

第 1 の計算機システムと、該第 1 の計算機システムに接続される第 1 のストレージシステムとを有する正システムと、第 2 の計算機システムと、該第 2 の計算機システムに接続された第 2 のストレージシステムとを有する副システムとを備え、かつ少なくとも前記第 1、第 2 のストレージシステムの間が相互接続されているシステムにおけるデータ二重化制御方法であって、

前記第 1 の計算機システムの処理によって前記第 1 のストレージシステムが保持するデータベースを更新するとともに該第 1 のストレージシステム内の特定記憶装置に前記データベースのログを登録するステップと、

前記ログの複製のために設定された第 2 のストレージシステム内の特定記憶装置に前記第 1 のストレージシステムの特定記憶装置のログの変更をコピーするステップと、

前記コピーステップにより前記第 2 のストレージシステム内の特定記憶装置の保持内容に変更が生じたことを前記第 2 の計算機システムに通知するステップと、

前記第 2 の計算機システムが前記第 2 のストレージシステム内の特定記憶装置の保持内容の変更を読み込むステップと、

前記第 2 の計算機システムが読み込んだログを実行して前記第 2 のストレージに保持す前記データベースの複製を更新するステップ、  
を有することを特徴とするデータ二重化制御方法。

【請求項 9】

前記第 2 のストレージシステム内の特定記憶装置の保持内容に変更が生じたことの前記第 2 の計算機システムへの通知は、一定時間毎に行うことを特徴とする

請求項 8 のデータ二重化制御方法。

【請求項 1 0】

前記第 2 のストレージシステム内の特定記憶装置の保持内容に変更が生じたことの前記第 2 の計算機システムへの通知は、前記第 1 のストレージシステムから前記第 2 のストレージシステムを制御するインターフェースを介して前記第 2 のストレージシステムを制御することにより行うことを特徴とする請求項 8 記載のデータ二重化制御方法。

【請求項 1 1】

前記第 2 のストレージシステムの制御は前記第 1 の計算機システムからの指示により行うことを特徴とする請求項 1 0 記載のデータ二重化制御方法。

【請求項 1 2】

請求項 8 記載のデータ二重化制御方法において、前記第 1 の計算機システムが停止したことを検知するステップと、前記第 2 の計算機システムが前記第 1 の計算機システムから業務を引き継ぐステップを更に有することを特徴とするデータ二重化制御方法。

【請求項 1 3】

請求項 1 2 記載のデータ二重化制御方法において、前記第 2 計算機システムが業務を引き継いだ後、前記第 1 の計算機システムを修復するステップと、前記正システムと副システムの相互関係を入れ替えて前記第 2 のストレージシステムの保持するデータベースの複製を前記第 1 のストレージシステムに形成するステップを更に有することを特徴とするデータ二重化制御方法。

【請求項 1 4】

第 1 の計算機システムと、該第 1 の計算機システムに接続される第 1 のストレージシステムとを有する正システムと、第 2 の計算機システムと、該第 2 の計算機システムに接続された第 2 のストレージシステムとを有する副システムとを備え、かつ少なくとも前記第 1、第 2 のストレージシステムの間が相互接続されているシステムにおけるデータ二重化制御方法であって、

前記第 1 の計算機システムの処理によって前記第 1 のストレージシステムが保持するデータベースを更新するとともに該第 1 のストレージシステム内の特定記

憶装置に前記データベースのログを登録するステップと、

前記ログの複製のために設定された第2のストレージシステム内の特定記憶装置に前記第1のストレージシステムの特定記憶装置のログの変更をコピーするステップと、

前記コピーステップにより前記第2のストレージシステム内の特定記憶装置の保持内容に変更が生じたことを前記第2の計算機システムが検知するステップと、

前記第2の計算機システムが前記第2のストレージシステム内の特定記憶装置の保持内容の変更を読み込むステップと、

前記第2の計算機システムが読み込んだログを実行して前記第2のストレージに保持する前記データベースの複製を更新するステップ、

を有することを特徴とするデータ二重化制御方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は計算機およびストレージ装置からなる業務システムに関し、特に複数のシステム間でデータを複製するデータ二重化制御方法、およびデータを複製したシステムへの高速な切り替えを実現する方法に関する。

【0002】

【従来の技術】

データベースサーバおよびストレージ装置からなる業務システムが複数ある場合のデータ複製方法として、データベースサーバ上で動作するDBMSが実行する方法がある。DBMSがデータ複製する方法については、例えば非特許文献1に記述がある。複数のシステムのデータベースサーバ同士を接続し、一方のシステム上で動作するDBMSの更新情報を別のシステムに転送することでデータ複製する技術である。

【0003】

また、同様なシステムのデータ複製方法として、ストレージ装置間のデータコピー機能を使用する方法がある。ストレージ装置間のデータコピー機能について



は、例えば非特許文献 2 に記述がある。複数システムのストレージ同士をファイバーチャネルで接続し、一方のストレージ装置のディスクドライブに更新があると、別のストレージ装置のディスクドライブにもデータの更新を反映する技術である。

【非特許文献 1】

Oracle9i製品カタログ (<http://www.oracle.co.jp/products/catalog/pdf/9iDBr2J07266-01.pdf>)、第 6 頁。

【非特許文献 2】

日立統合ストレージソリューション「Storeplaza」カタログ(<http://www.hitachi.co.jp/Prod/comp/storeplaza/data/stpzclg.pdf>)、第 5 頁。

【 0 0 0 4】

【発明が解決しようとする課題】

従来のデータ複製方法を実行する場合、通常のデータベース業務を行う以上のコストが必要である。また、複数システム間で同期したデータ複製を行うと業務処理の遅延が発生する。

【 0 0 0 5】

DBMS によるデータ複製を行うためには、DBMS の動作するデータベースサーバが業務処理を行う負荷に加えて、データ複製処理を行う負荷が加わるためにより高性能なデータベースサーバが必要でありコストが増加するという課題がある。また、複製したデータが一致するためには、DBMS の更新処理を実行するたびにデータベースサーバ間で同期通信を行う必要がある。同期通信中は、DBMS が次の更新処理を実行できないため、業務が遅延することが課題である。

【 0 0 0 6】

ストレージ装置でデータ複製を行うためには、DBMS が扱うデータの更新を全てコピーするため、ストレージ装置間の接続に広帯域の回線を使う必要がある。広帯域の回線を使用することでコストが増大する課題がある。また、複製したデータが一致するためには、ディスクドライブ上のデータが更新されるたびにストレージ装置間で同期通信を行う必要がある。同期通信中はディスクドライブへの次の更新処理が実行できないため、業務が遅延することが課題である。



## 【 0 0 0 7 】

また、同期通信による遅延を防ぐため、DBMSやストレージ装置間の通信を非同期で実行する方法が存在するが、障害や災害でデータ複製先のシステムに切り替える場合に、未転送分のデータを複製先で再構築する必要が生じ、システムの切り替えが遅延することが課題である。

## 【 0 0 0 8 】

## 【課題を解決するための手段】

サーバが外部から受け付けた要求に応じて業務を実行すると、ストレージ装置に保存されたデータへの更新や追加が必要となる。このストレージ装置のデータ複製を行うために全てのデータを複製するのではなく、複製先としてサーバとストレージ装置を用意し、複製元のサーバで実行された業務を復元可能なログをストレージ装置の特定のディスクドライブに保存し、このディスクドライブが更新されるたびに複製先のストレージ装置にディスクドライブのコピーを行う。複製先のストレージ装置へのディスクドライブのコピーが完了したら、コピーされたストレージ装置からログを保存したディスクドライブが更新されたことを複製先のサーバに通知する。複製先のサーバは、ストレージ装置からログを保存したディスクドライブの変更通知を受信できるようにしておき、通知を受けた後でログをディスクドライブから読み取り、複製元のサーバで行われたのと同じ業務処理を実行する。このログを基にした業務処理の実行後、その結果をストレージ装置に反映することでデータの複製が完了する。

## 【 0 0 0 9 】

このデータ複製方法を実行しているシステムにおいて、複製元のサーバとストレージ装置が障害や保守操作により停止した場合、複製先のストレージ装置に保存された業務データが最新の状態にあるため、複製元のサーバが受信していた業務を複製先で受信するように変更することで、業務の処理を中止せずにサーバとストレージ装置の切り替えを実行する。

## 【 0 0 1 0 】

サーバとストレージ装置の切り替え実行後に、複製元と複製先双方のサーバとストレージ装置とがデータ複製のために実行していた処理を交替することで、業

務を受信し処理しているシステムが停止した場合、再びサーバとストレージ装置の切り替えを実行する。

【 0 0 1 1 】

【発明の実施の形態】

以下で説明する実施形態では、例として計算機上で動作する業務としてデータベースサーバを取り上げるが、計算機上で実行される業務はデータベースに限定するものではない。

〔第 1 実施形態〕

図 1 は、本発明が適用されたデータベースサーバとストレージ装置を用いたデータ複製システムの一実施例である。

【 0 0 1 2 】

正システムを構成するのは、データベースサーバ 2 とストレージ装置 8 である。これらはデータベースサーバ 2 に内蔵されたストレージ接続装置 3 とストレージ装置 8 のディスク制御装置 5 とがサーバ・ストレージ間接続インタフェース 4 によって接続される。ストレージ装置 8 はディスク制御装置 5 によって読み込み書き込みを行うデータを保存するディスクドライブ 6, 7 を内蔵しており、データベースサーバ 2 が業務ネットワーク 1 を通じて業務要求を受け取って処理したデータや、その処理に必要なデータおよびデータベースサーバ 2 内部で実行された業務データを保持する。

【 0 0 1 3 】

データベースサーバ 2 とストレージ装置 8 とは、サーバ・ストレージ間接続インタフェース 4 を通してデータの読み込み・書き込みを行うだけでなく、データベースサーバ 2 が要求したディスクドライブ 6, 7 の変更があった場合、ストレージ装置 8 からデータベースサーバ 2 に通知を行う方法を有している。

【 0 0 1 4 】

副システムを構成するのは、データベースサーバ 1 2 とストレージ装置 1 8 である。これらはデータベースサーバ 1 2 に内蔵されたストレージ接続装置 1 3 とストレージ装置 1 8 のディスク制御装置 1 5 とがサーバ・ストレージ間接続インタフェース 1 4 によって接続される。ストレージ装置 1 8 はディスク制御装置 1

5によって読み込み書き込みを行うデータを保存するディスクドライブ16, 17を内蔵しており、データベースサーバ12が業務ネットワーク1を通じて業務要求を受け取って処理したデータや、その処理に必要なデータおよびデータベースサーバ12内部で実行された業務データを保持する。

【0015】

データベースサーバ12とストレージ装置18とは、サーバ・ストレージ間接続インタフェース14を通してデータの読み込み・書き込みを行うだけでなく、データベースサーバ12が要求したディスクドライブ16, 17の変更があった場合、ストレージ装置18からデータベースサーバ12に通知を行う方法を有している。

【0016】

ディスク制御装置5とディスク制御装置15とはストレージ装置間接続インタフェース20により接続される。これにより、正システムのストレージ装置8と副システムのストレージ装置18は互いに接続される。ストレージ装置8とストレージ装置18は、一方のディスクドライブの一つを複製元に、他方のディスクドライブの一つを複製先にあらかじめ設定しておくことで、ストレージ装置間接続インタフェース20を通して内容を複製する方法を有している。

【0017】

以下、本実施形態のデータ複製方法およびシステム切り替え方法の動作を説明する。本実施形態では業務を通常実行している正システムと、正システムが何らかの理由で稼働不可能になった時に業務を引き継ぐ副システムとの間でデータ複製を行うものとする。

【0018】

まず、データ複製方法を実現するための初期設定を正システム、副システム双方について行う。

【0019】

正システムの初期設定は、業務システムに応じたデータベースを構築することから始める。ストレージ装置8のディスク制御装置5で、データベースサーバ2が使用可能なディスクドライブ6, 7を割り当てる。データベースサーバ2はデ

データベースのデータを保持するディスクドライブ6とデータベースのログを保持するディスクドライブ7とを設定する。ここで言うログとは、データベースの更新作業を逐一表すもので、ログを再実行することでデータベースの再構築が可能なものである。例えば、データベースが実行したトランザクションログやデータベースサーバが受け取った業務要求全てのSQLである。

## 【 0 0 2 0 】

副システムにも正システムと同様のデータベースを構築する。ストレージ装置18において、ストレージ装置8でデータベースサーバ2が使用可能としたディスクドライブ6と同様なディスクドライブ16と、ディスクドライブ7と同様なディスクドライブ17をディスク制御装置15でデータベースサーバ12が使用可能となるように割り当てる。データベースサーバ12は、データベースサーバ2同様に、データベースのデータを保持するディスクドライブ16とデータベースのログを保持するディスクドライブ17とを設定する。

## 【 0 0 2 1 】

次に、正システムのストレージ装置8と副システムのストレージ装置18との間で、ストレージ装置間接続インタフェース20を通じてデータベースのログを保持するディスクドライブ7をディスクドライブ17にコピーするように設定する。このディスクドライブコピーは、同期コピー、非同期コピーいずれとも可能であるが、非同期コピーを行う場合、正システムと副システムのログディスクが常に一致するとは限らず、システム切り替え時にデータが欠損することがある。

## 【 0 0 2 2 】

そして、副システムのデータベースサーバ12からストレージ装置18のディスク制御装置15にたいして、ディスクドライブ17の更新が行われたらデータベースサーバ12に通知を行うように設定する。

## 【 0 0 2 3 】

正システムに障害・災害などが発生した場合に、副システムに切り替えるため正システムの停止を迅速に検知する必要がある。そのため、正システムのデータベースサーバ2と副システムのデータベースサーバ12の間で正システムが稼働していることを通知するための通信設定を行う。例えば、正システムのデータバ

ースサーバ2から副システムのデータベースサーバ12に業務ネットワーク1を経由して一定時間間隔で通知を行う方法がある。また、正システムの稼働状態を監視する外部のサーバから副システムへの切り替えを指示する方法や、副システムから一定時間間隔で正システムに稼働状態を問い合わせる方法もある。

#### 【0024】

以上のような設定がデータベースサーバ2、12とストレージ装置8、18で完了した後、正システムのデータベースサーバ2で業務処理を開始する。以下では、データ複製の手順について説明する。

#### 【0025】

データ複製第1ステップ101：業務処理要求は、業務ネットワーク1を通じてデータベースサーバ2に到着する。業務処理要求は業務ネットワーク上のプロトコルに応じて送付され、データベースサーバ2の管理するデータ内容を参照するものや更新するものからなる。例えば、TCP/IPプロトコルによって送付される、SQLコマンドの組み合わせからなる。

#### 【0026】

業務処理要求を受信したデータベースサーバ2は、ネットワークプロトコル層の解析を行い、データベースへの業務処理内容を取り出し、業務処理内容の解析を行った後、業務処理を実行する。例えば、TCP/IPプロトコルの解析を行い、SQLコマンドを取り出し、その処理をデータベースで実行する処理がある。

#### 【0027】

データ複製第2ステップ102：業務処理の内容が、データベースの更新処理を伴う場合にはストレージ装置内に保持しているデータを更新する必要がある。その場合には、ストレージ装置接続装置3からサーバ・ストレージ接続インタフェース4を通じて、ディスク制御装置5に対してディスクドライブ6へのデータ更新とディスクドライブ7への更新ログの書き込みをストレージ装置8に指示する。例えば、データベースサーバ2にホストバスアダプタを装着し、ファイバーチャネルケーブルを通じてSCSIコマンドをディスクコントローラに送信することに当たる。また、本実施形態では簡単のため1回の書き込み要求のように図



示したが、通常は別のディスクドライブへの書き込み要求は複数の要求に分けて送信される。

## 【 0 0 2 8 】

データ複製第3ステップ103：データ書き込みの要求を受けたディスク制御装置5は、ディスクドライブ6へデータ更新の書き込みを行う。

## 【 0 0 2 9 】

データ複製第4ステップ104：また、ディスク制御装置5はディスクドライブ7へ更新ログの書き込みを行う。

## 【 0 0 3 0 】

データ複製第5ステップ105：ディスクドライブ7はそのデータの更新書き込みが終了すると副システムのストレージ装置18内のディスクドライブ17にコピーするように設定されているため、ディスク制御装置5はストレージ装置間接続インタフェース20を通じて副システムのストレージ装置18にあるディスク制御装置15にディスクドライブ7の更新内容を送信し、ディスクドライブ17に書き込むよう指示し、ディスク制御装置15はディスクドライブ17へと書き込みを行う。例えば、ストレージ装置間接続インタフェース20としてファイバーチャネルケーブルを用い、ストレージ装置の管理ソフトウェアでディスクドライブ7、17のコピーを設定することで実現できる。また、本実施形態では、ディスクドライブ7の更新直後にディスクドライブ17へのコピーを行う同期コピー方法としているが、一定時間間隔でコピーを実行する非同期コピー方法を用いることも可能である。ただし、非同期コピー方法を用いた場合、システム切り替え時にディスクドライブのデータがコピーされていない事態も発生しうる。

## 【 0 0 3 1 】

データ複製第6ステップ106：ディスクドライブ17への更新を実行後、ディスク制御装置15はあらかじめデータベースサーバ12から更新を通知するように指定されているため、更新が発生したことをデータベースサーバ12に通知する。この更新通知要求と更新通知のインタフェースは、例えば、データベースサーバ12からストレージ装置18内の特殊なディスクドライブへの読み込み要求の応答としてディスク制御装置15が通知する方法や、データベースサーバ1

2 から更新通知を要求するディスクドライブ 1 7 への専用コマンドに対する応答としてディスク制御装置 1 5 が通知する方法、また、ディスク制御装置 1 5 からディスクの更新を通知する専用の割り込みインタフェースをデータベースサーバ 1 2 内のストレージ接続装置 1 3 に設ける方法などがある。また、ディスク制御装置 1 5 からデータベースサーバ 1 2 への通知は、更新が発生する度に実行する方法に限定するわけではなく、一定時間間隔ごとに通知する方法や、データベースサーバ 2 から指示をストレージ装置 8 に発行したものをストレージ装置 1 8 に伝えてデータベースサーバ 1 2 への通知を実行させる方法などがある。

## 【 0 0 3 2 】

データ複製第 7 ステップ 1 0 7 : ディスクドライブ 1 7 の更新通知を受けたデータベースサーバ 1 2 は、ディスクドライブ 1 7 の更新分を読み込み、そこに書き込まれた更新ログを実行し、ディスクドライブ 1 6 上のデータを更新するようにストレージ接続装置 1 3 からサーバ・ストレージ間接続インタフェース 1 4 を通じてディスク制御装置 1 5 に通知する。例えば、データベースサーバ 1 2 にホストバスアダプタを装着し、ファイバーチャネルケーブルを通じて S C S I コマンドをディスクコントローラに送信する方法がある。

## 【 0 0 3 3 】

正システムのデータベースサーバ 2 が業務処理要求を実行するたびに、このように正システムのデータを副システムに複製することで、正システムのデータベースサーバ 1 2 にデータ複製のための負荷をかけることなく、また業務ネットワーク 1 上にデータ複製のためのデータ送信を行うことなく、ストレージ装置 8 , 1 8 間でのデータ転送量を小さくして、データ複製のコストを抑え、業務の遅延を小さくすることができる。

## 【 0 0 3 4 】

正システムが災害や機器の障害などにより停止した場合、副システムに業務処理を切り替える。正システムの保守作業を行うため必要がある場合でも、正システムを停止させ、副システムに業務処理を切り替えることがある。図 2 は正システム停止後に副システムが業務処理を引き継ぐ処理を行う手順を示した。装置構成は図 1 と同様であるため詳細は省略する。以下では、業務処理引き継ぎの手順



を説明する。

【0035】

正システムが停止すると、システムを切り替えて副システムのデータベースサーバ12とストレージ装置18で業務処理を引き継ぐ。正システムの停止は、例えばデータベースサーバ2とデータベースサーバ12との間で一定時間間隔で通信を行うハートビート通信や、データベースサーバ2、12以外の監視サーバを業務ネットワーク1に接続してハートビート通信を行う方法で検出可能である。

【0036】

システム切り替え第1ステップ201：まず、データベースサーバ12がログの書き込まれたディスクドライブ17を参照し、未実行の業務がディスクドライブ17に存在するか確認する。

【0037】

システム切り替え第2ステップ202：未実行の業務処理があればその業務を実行してディスクドライブ16のデータ更新を行うようストレージ接続装置13からサーバ・ストレージ間接続インタフェース14を通じてディスク制御装置15に通知する。

【0038】

システム切り替え第3ステップ203：ディスク制御装置15は、要求を受けたデータの書き込みをディスクドライブ16に行う。

【0039】

ディスクドライブ16のデータ更新が完了したら、データベースサーバ2で受け取っていた業務処理要求をデータベースサーバ12が受け取れるように設定を変更する。例えば、データベースサーバ2が業務要求受信に用いていたネットワークアドレスを引き継ぐ方法がある。

【0040】

システム切り替え第4ステップ204：その後、データベースサーバ12で業務要求を受け付けて業務処理を開始する。

【0041】

さらに、正システムが障害・災害から回復し再び動作するようになった場合や

保守作業完了で正システムが動作可能になった場合、本実施形態で説明してきたデータ複製方法を、副システムから正システムに複製する方向に適用することで、正システムが停止中に副システムで実行した業務処理によるデータやログの更新を正システムに反映させることができる。

【 0 0 4 2 】

例えば、ストレージ装置 8，18 の間でディスクドライブ 17 の更新部分をディスクドライブ 7 にコピーする設定を行い、データベースサーバ 2 でディスクドライブ 7 の更新ログを実行することで、副システムのデータ複製を実行可能である。このように、正システムと副システムが同時に停止することがなければ、交互に本発明のデータ複製方法を適用することで業務停止時間を小さくできる。

【 0 0 4 3 】

また、本実施形態では正システムと副システムが一对一の形態を説明したが、正システムから複数の副システムへのデータ複製を行う方法や、正システムから副システムへ複製したデータをさらに別の副システムへデータ複製を行なう方法も容易に構築可能である。

【 0 0 4 4 】

図 3 から図 6 に、本実施形態の主な構成要素であるデータベースサーバおよびストレージ装置の処理手順をフローチャートで示した。以下で各図のフローチャートについて説明する。

【 0 0 4 5 】

図 3 に示した正システムデータベースサーバ処理手順のフローチャートについて説明する。

【 0 0 4 6 】

まず、データベースサーバの初期設定を行なう（301）。例えば、初期設定には、データベースの構築やディスクドライブの割り当てなどがある。

【 0 0 4 7 】

次に、データ複製システムを構築するまで、ストレージ装置と副システムの初期設定完了を待つ（302）。例えば、ストレージ装置間のディスクコピーの設定や副システムのデータベース構築の完了を待つことになる。

【 0 0 4 8 】

データ複製システムの初期設定が一通り完了すると、業務処理要求受付を開始する（3 0 3）。例えば、インターネット経由で行う商取引の商品管理などが業務処理にあたる。

【 0 0 4 9 】

データベースとして稼動を開始したら、副システムに稼動状態を通知する時刻か（3 0 4）定期的に判定を行なう。自身が稼動中であることを副システムに知らせ、システム切り替えが必要かを通知するためである。

【 0 0 5 0 】

もし通知する時刻であれば、副システムに稼動状態を通知（3 1 0）し、自身で管理している稼動状態の通知時刻を更新（3 1 1）する。もし通知する時刻でなければ、業務処理要求が到着したか（3 0 5）判定する。

【 0 0 5 1 】

業務処理要求が到着していれば、業務処理を実行する（3 0 6）。業務処理がデータ更新を伴う場合、データの更新をストレージ装置に送る（3 0 7）。そして、行なった業務処理のログをストレージ装置に書き込む（3 0 8）。

【 0 0 5 2 】

ストレージ装置に対して行なった書き込み要求について、ストレージ装置からの書き込み完了報告を受信する（3 0 9）ことで業務処理要求が完了する。

【 0 0 5 3 】

正システムは一度稼動すると、ここで説明したように、副システムへの稼動状態の通知と業務処理要求の実行を繰り返し行なう。

【 0 0 5 4 】

図 4 に示した副システムデータベースサーバ処理手順のフローチャートについて説明する。

【 0 0 5 5 】

正システム停止後に交替して業務処理を実行するのがデータ複製の目的であるため、副システムのデータベースサーバには正システムの設定にあわせた初期設定を行う（4 0 1）。例えば、コピーするログディスクやデータディスクの用意

などがデータ複製のためには必要となる。

【0056】

次に、正システムからコピーされたログを参照するため、ストレージ装置にログディスクの更新通知を指示する（402）。これにより、ログの更新を検出できる。

【0057】

そして、ログが更新されるとその処理を副システムで実行してデータベースのデータも更新する、データ複製処理を開始する（403）。

【0058】

システム切り替えが必要かを判定するため、正システムから稼動状態通知があったか（404）確認する。例えば、10秒間通知がない場合にシステムを切り替えるというように方針決めておき、判定を実行することになる。

【0059】

もし通知がない場合には、システム切り替えの処理を行なう。まず、ログディスクの更新分で未実行の業務処理を実行する（410）。そして実行の結果、データの更新をストレージ装置に送る（411）。ストレージ装置のデータ更新完了報告を受信する（412）ことで、正システムのデータ複製が完了したとみなす。そして、正システムの業務引継処理を実行する（413）ことで、業務ネットワークと接続可能としてから、業務処理要求受付を開始する（414）。

【0060】

もし稼動状態通知があれば、ストレージ装置からログディスク更新通知があったか（405）判定する。ログディスク更新通知があれば、ログディスクの更新分を読み込む（406）。そして、その更新分について業務処理を実行する（407）。業務処理実行によって発生するデータの更新をストレージ装置に送る（408）ことで、データが正システムの最新のものと一致するようにする。ストレージ装置のデータ更新完了報告を受信する（409）と再び状態通知の受信やストレージ装置の更新通知待ちの処理を繰り返してデータ複製をしながら、システム切り替えの準備をする。

【0061】

図 5 に示した正システムストレージ装置処理手順のフローチャートについて説明する。

【 0 0 6 2 】

まず、ストレージ装置内のディスクドライブをデータベースサーバに割り当てるなどの、初期設定を行う（5 0 1）。

【 0 0 6 3 】

そして、本実施形態のデータ複製方法を行なうため、正システムのログディスクを副システムのログディスクに対応付けしコピーの設定をする（5 0 2）。この設定を行なう前に、副システムのデータベースサーバとストレージ装置の初期設定を完了しておく必要がある。

【 0 0 6 4 】

設定が完了したら、読み込み・書き込み処理を開始（5 0 3）し、データベースサーバからのデータ更新要求などを受け付ける状態になる。

【 0 0 6 5 】

処理要求受信（5 0 4）を待つ状態から要求を受信すると、まず書き込み要求か（5 0 5）判定する。書き込み要求でなければ、要求された読み込みデータをデータベースサーバに転送（5 1 1）し、データベースサーバにデータ読み込み完了報告を送信する（5 1 2）。実際はディスクドライブのコントロールなどの要求も受信するが、ここでは読み込み要求と同じものとみなしている。書き込み要求を受信した場合には、ディスクにデータを書き込む（5 0 6）処理を行い、そのディスクがコピーを設定したディスクか（5 0 7）判定する。コピー設定されていなければ、データベースサーバにデータ書き込み完了報告を送信する（5 1 0）。コピーを設定したディスクであれば、副システムのストレージ装置に書き込みデータを転送（5 0 8）し、副システムのストレージ装置から書き込み完了報告を待つ（5 0 9）。副システムから完了報告を受け取るとデータベースサーバにデータ書き込み完了報告を送信する（5 1 0）。ここでは、正システムと副システムのストレージ装置間で同期コピーを行なう方法としている。

【 0 0 6 6 】

このように、正システムのストレージ装置はデータベースサーバからの処理要

求を待ち、ディスクのデータ転送、副システムへのディスクコピー処理を繰り返す。

【 0 0 6 7 】

図 6 に示した副システムストレージ装置処理手順のフローチャートについて説明する。

【 0 0 6 8 】

まず、ストレージ装置内のディスクドライブをデータベースサーバに割り当てや外部ストレージ装置からのディスクコピー設定などの、初期設定を行う（6 0 1）。

【 0 0 6 9 】

そして、読み込み・書き込み処理開始（6 0 2）後、データベースサーバからの要求を受信可能な状態となる。さらに、データベースサーバから更新通知するディスクを指定される（6 0 3）ことでデータ複製の準備が整う。

【 0 0 7 0 】

処理要求受信（6 0 4）を開始し、処理要求を受けるとその要求が書き込み要求か（6 0 5）判定する。書き込み要求でなければ、要求されたデータを要求元に転送（6 1 0）し、要求元にデータ転送完了報告を送信する（6 1 1）ことで処理要求の実行が完了する。一方、書き込み要求であった場合、まず、ディスクにデータを書き込む（6 0 6）。そして、要求元にデータ書き込み完了報告を送信する（6 0 7）。通常の処理要求はこれで処理が完了するが、データ複製方法を実施するためにディスクへの書き込みが発生した場合は、そのディスクがデータベースサーバに更新通知を指定されたディスクか（6 0 8）判定し、指定されたディスクであれば、データベースサーバに更新を通知する（6 0 9）。指定されていない場合は、通知せずに処理を完了する。

【 0 0 7 1 】

このように、副システムのストレージ装置は正システムからのログディスク書き込みと副システムのデータベースサーバからの読み込み・書き込み要求を処理し、更新を通知するディスクであればその通知を送信する処理を繰り返して、データ複製処理を実現する。



【 0 0 7 2 】

本実施形態のように正システムから副システムに切り替えるが、上記のようなデータ複製方法を用い、特にストレージ装置 8，18 間のディスクコピーが同期コピー方法で行うとシステム切り替えに伴う遅延を小さくすることができる。

〔第 2 実施形態〕

第 1 実施形態では、ディスクドライブの更新をデータベースサーバに通知するために通常のディスクドライブ読み込み書き込み以外のインタフェースを必要としたが、以下で説明する第 2 実施形態においてはディスクドライブの読み込み書き込みインタフェースのみでデータ複製を実現する。データ複製システムの構成は、図 1 に示す第 1 実施形態と同様に構成される。

【 0 0 7 3 】

第 1 実施形態では、副システムにおいてディスクドライブ 17 の更新をデータベースサーバ 12 に通知するための設定をした。これに対し、本実施形態ではデータベースサーバ 12 からディスクドライブ 17 をポーリングで監視し、更新を検知する。

【 0 0 7 4 】

更新の検知は以下のような手順で可能である。データベースサーバ 12 で、ディスクドライブ 17 にログが書き込まれる位置を保持し、その位置のデータを定期的に読み込んで更新されたかを判定する。更新されていれば、ログに従って業務処理をデータベースサーバ 12で行う。処理が完了したら、ログが書き込まれる位置の更新を行い、再び定期的に読み込んで更新されたかの判定処理を繰り返す。

【 0 0 7 5 】

データベースの更新ログは、通常一定の領域に順次上書きされないようにディスクに書き込む。そして、領域の終端まで書き込むと再び領域の先頭から順次書き込む。そのため、更新ログの書き込みが上書きされる前にデータベースサーバで内容を読み込むことができることと、更新ログを 1 つずつ区別することができるようになっていたことが保証される場合、ポーリングで監視してデータのディスクドライブを更新することでデータの複製が可能である。



## 【 0 0 7 6 】

このように、データベースサーバからディスクドライブの更新をポーリングによって監視する方法でデータ複製する場合、ポーリング間隔を十分小さくすることによって、システム切り替えによる遅延を小さくすることができる。また、第 1 実施形態と同様のシステム構成であり、データ複製にかかるコストを小さくできる。

## 〔第 3 実施形態〕

第 1 実施形態、第 2 実施形態では、データベースサーバ 2， 1 2 とストレージ装置 8， 1 8 が直接接続された場合やストレージエリアネットワークで接続されたことを前提としていたが、本実施形態では、ストレージ装置としてネットワークアタッチトストレージ（N A S）装置を使用して実現する。本実施形態のデータ複製システムの構成は、図 1 に示す第 1 実施形態と同様の構成である。

## 【 0 0 7 7 】

本実施形態では、ストレージ装置間のディスクコピー方法が第 1 実施形態、第 2 実施形態と異なる。N A S 装置はファイルシステムでのアクセス要求を受信する。そのため、ディスク制御装置 5， 1 5 間のストレージ間接続インタフェースもファイル単位でのアクセスを実行する。そのため、ディスク制御装置 5 内でディスクドライブ 7 の変更を検知するのではなく、データベースサーバ 2 が操作するログファイルの更新を検知する必要がある。更新の検知には、ディスク制御装置 5 でログファイルの更新を定期的に監視するデーモンを実行しておき、更新が起こったらストレージ間接続インタフェースを通じてファイルのコピーをディスクドライブ 1 7 に書き込む。また、副システムでログの更新を通知するインタフェースも、前述の第 1 実施例、第 2 実施例のものと異なる。ディスク制御装置 1 5 にはログファイル更新を検知するデーモンを備える。データベースサーバ 1 2 にはこのデーモンと通信を行うプロセスを生成しておき、つまり、ログファイルの更新があったら通知される機構を構築する。あるいは、第 2 実施形態で示したデータベースサーバ 1 2 からポーリングで監視する方法をログファイルに適用してもよい。

## 【 0 0 7 8 】

また、ログデータのコピー方法としてディスク制御装置 5 と別のディスクドライブ単位で変更を検知可能なディスク制御装置をストレージ装置 8, 18 に設け、その間でディスクの更新を行う方法も可能である。この場合、別途設けたディスク制御装置により、第 1 実施形態のようにディスクドライブ 7 からディスクドライブ 17 へのデータコピーを実行する。ログデータの更新をデータベースサーバ 12 に通知する方法は、上記のようなディスク制御装置 15 のデーモンとデータベースサーバ 12 のプロセスで通信を行う方法やデータベースサーバ 12 からポーリングで監視する方法によって可能である。

## 【 0 0 7 9 】

また、NAS 装置ではログファイルをコピーする際に、ファイルを全てストレージ装置 8, 18 間で転送する必要がある。データ転送量を削減するため、ログをおくためのディレクトリを作成し、更新ログを 1 個ずつファイルとしてそのディレクトリに置いていくことで更新ログの転送量を削減可能である。データベースは更新ログを作成した日時をファイル名に使用して書き込む。これによりログの一意の識別が可能になる。また、一定以上の時間が経過したログを削除することで、ログの複製を保証することができ、ディスクドライブを使い尽くすことがなくなる。ログデータの更新のデータベースサーバ 12 への通知は、上記同様であるが、ディスク制御装置 15 で実行するログ更新を監視するデーモンは、ログディレクトリの下にあるファイルの監視を行い、新たなログファイルが作成された場合にデータベースサーバ 12 のプロセスに通知を行う。データベースサーバ 12 でファイル更新をポーリングで監視するデーモンを実行する場合も、同様にログディレクトリの下に新たなファイルが作成されたかを監視する。

## 【 0 0 8 0 】

このようにシステムを構築することで、データ複製が実現される。システム構成は第 1 実施形態と同様であり、コストを削減することが可能である。また、システム切り替え時の遅延も小さくすることが可能である。

## 【 0 0 8 1 】

## 【発明の効果】

本発明によれば、データベースサーバとストレージ装置からなる複数のシステ

ム間で、低コストかつ通常業務の遅延が小さいデータ複製を実現する。また、システム切り替え時の遅延を小さくすることができる。

【図面の簡単な説明】

【図 1】

正システムと副システムとの間でデータ複製システムおよびデータ複製方法の概念図である。

【図 2】

正システムと副システムとの間でデータ複製システムおよびシステム切り替え方法の概念図である。

【図 3】

正システムのデータベースサーバが行う処理手順のフローチャートである。

【図 4】

副システムのデータベースサーバが行う処理手順のフローチャートである。

【図 5】

正システムのストレージ装置が行う処理手順のフローチャートである。

【図 6】

副システムのストレージ装置が行う処理手順のフローチャートである。

【符号の説明】

- 1 : 業務ネットワーク
- 2 : 正システムのデータベースサーバ
- 3 : ストレージ接続装置
- 4 : サーバ・ストレージ間接続インタフェース
- 5 : ディスク制御装置
- 6 : データディスクドライブ
- 7 : ログディスクドライブ
- 8 : 正システムのストレージ装置
- 1 2 : 副システムのデータベースサーバ
- 1 3 : ストレージ接続装置
- 1 4 : サーバ・ストレージ間接続インタフェース

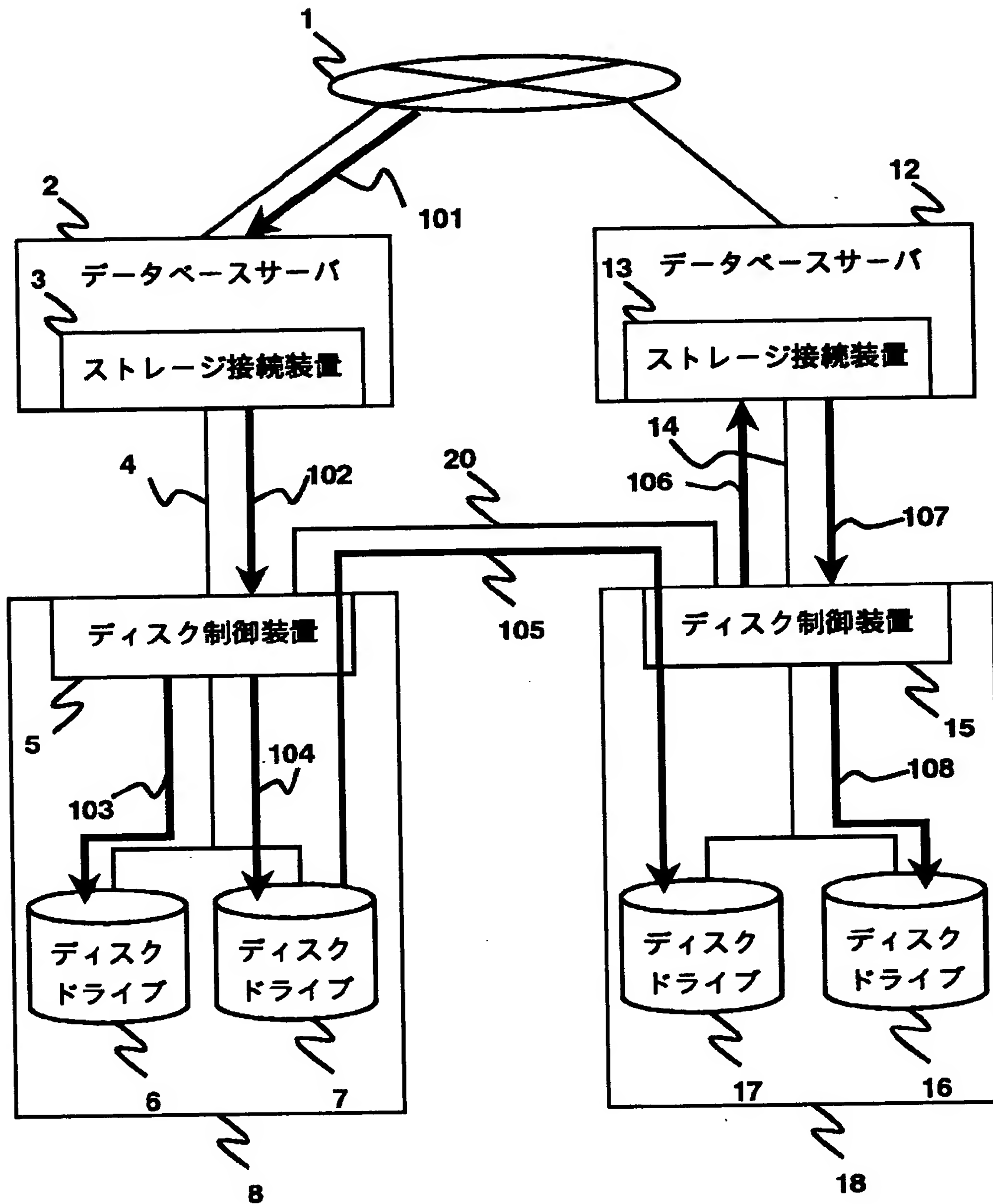
- 1 5 : ディスク制御装置
- 1 6 : データディスクドライブ
- 1 7 : ログディスクドライブ
- 1 8 : 副システムのストレージ装置
- 2 0 : ストレージ装置間接続インタフェース。

【書類名】

図面

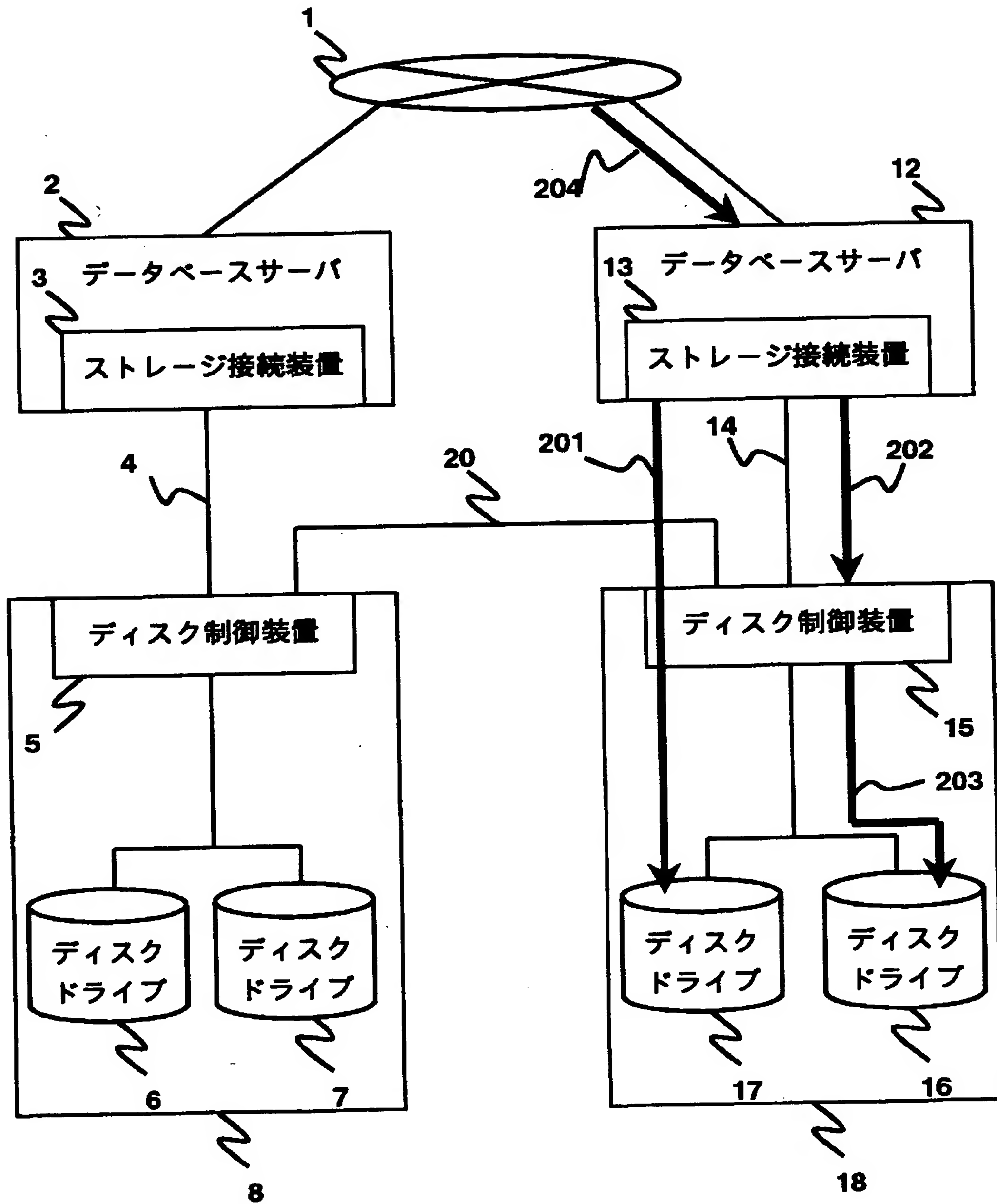
【図 1】

図1



【図2】

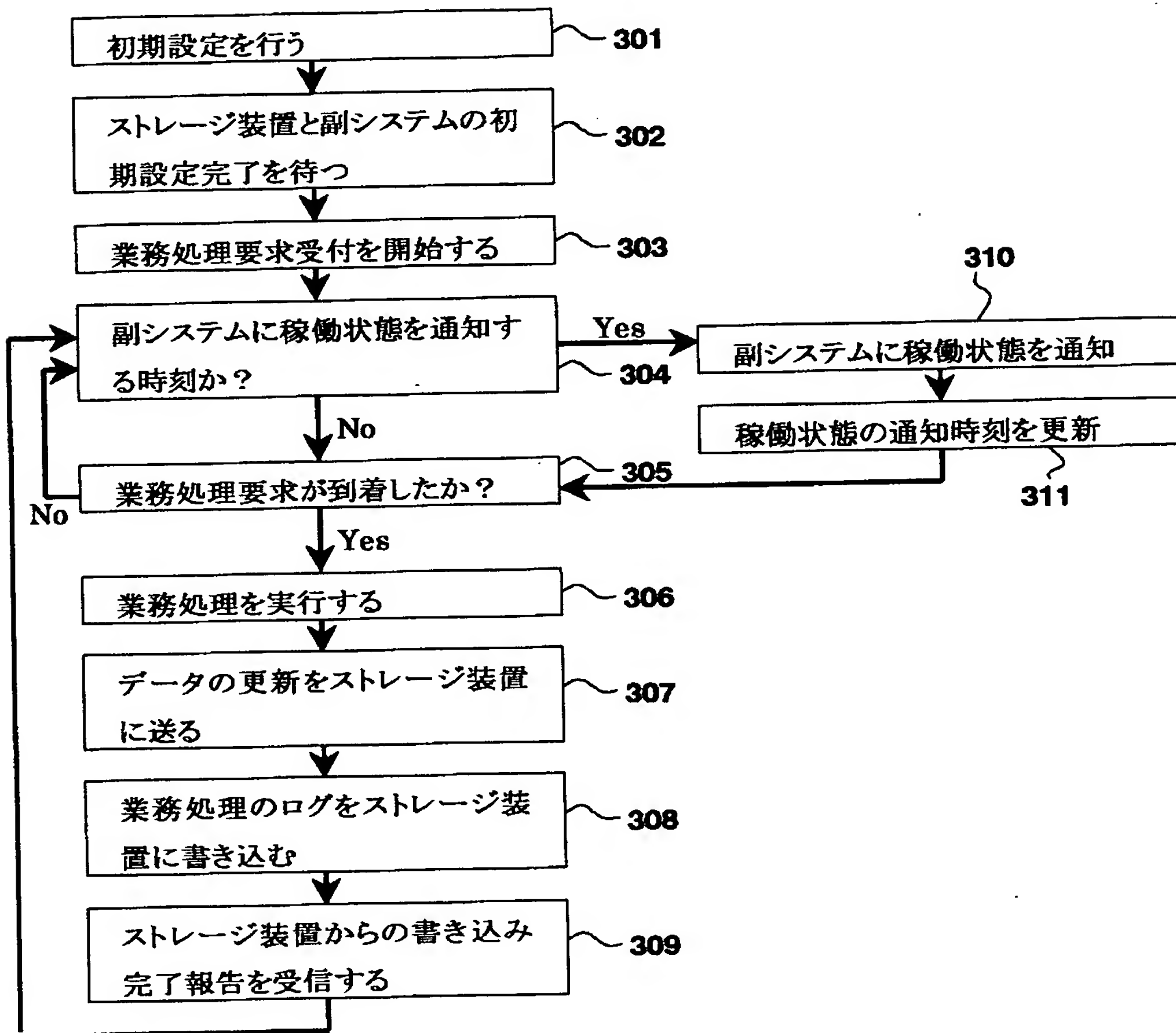
図2



【図 3】

図3

正システムデータベースサーバ処理手順

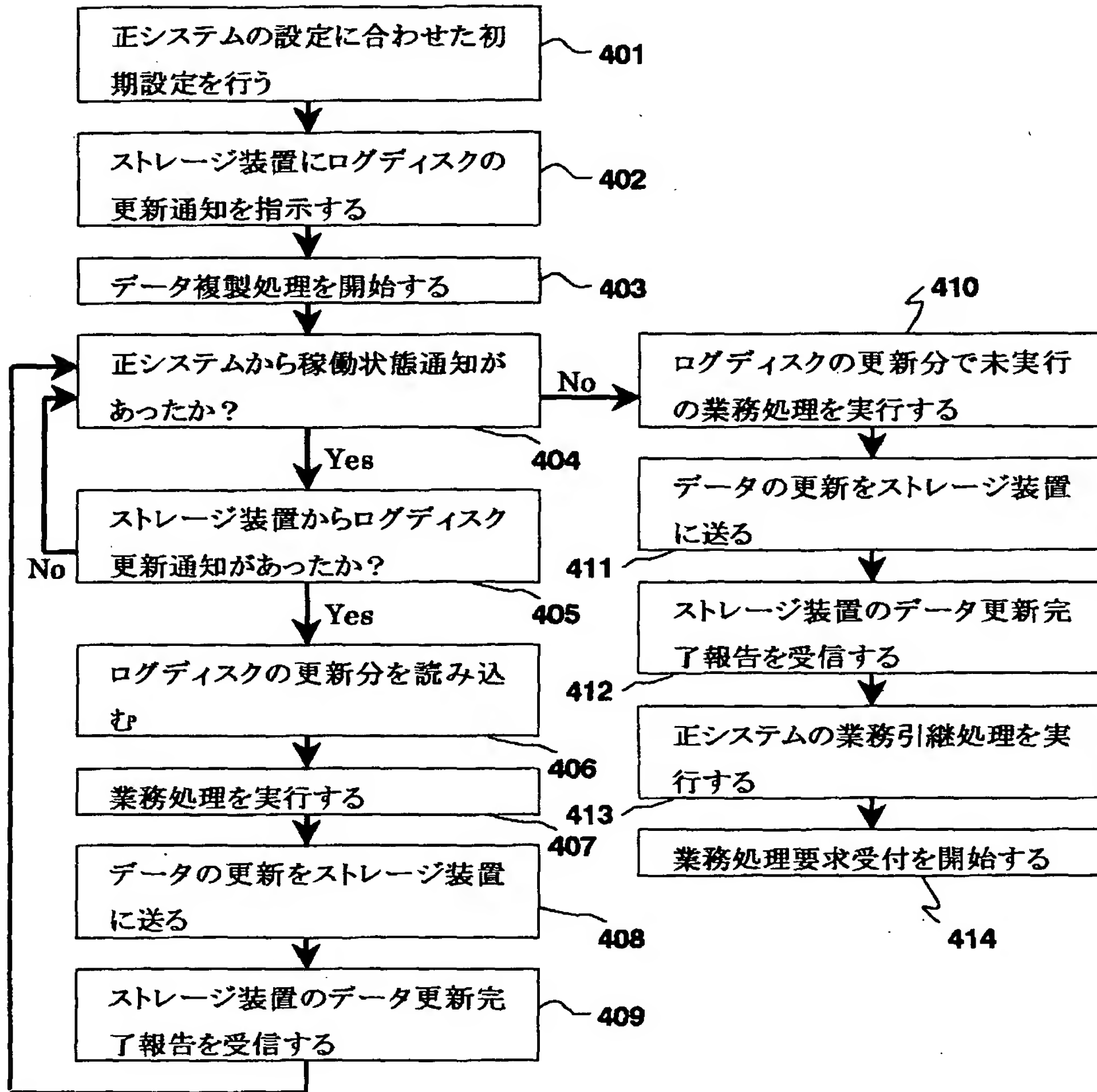




【図 4】

図4

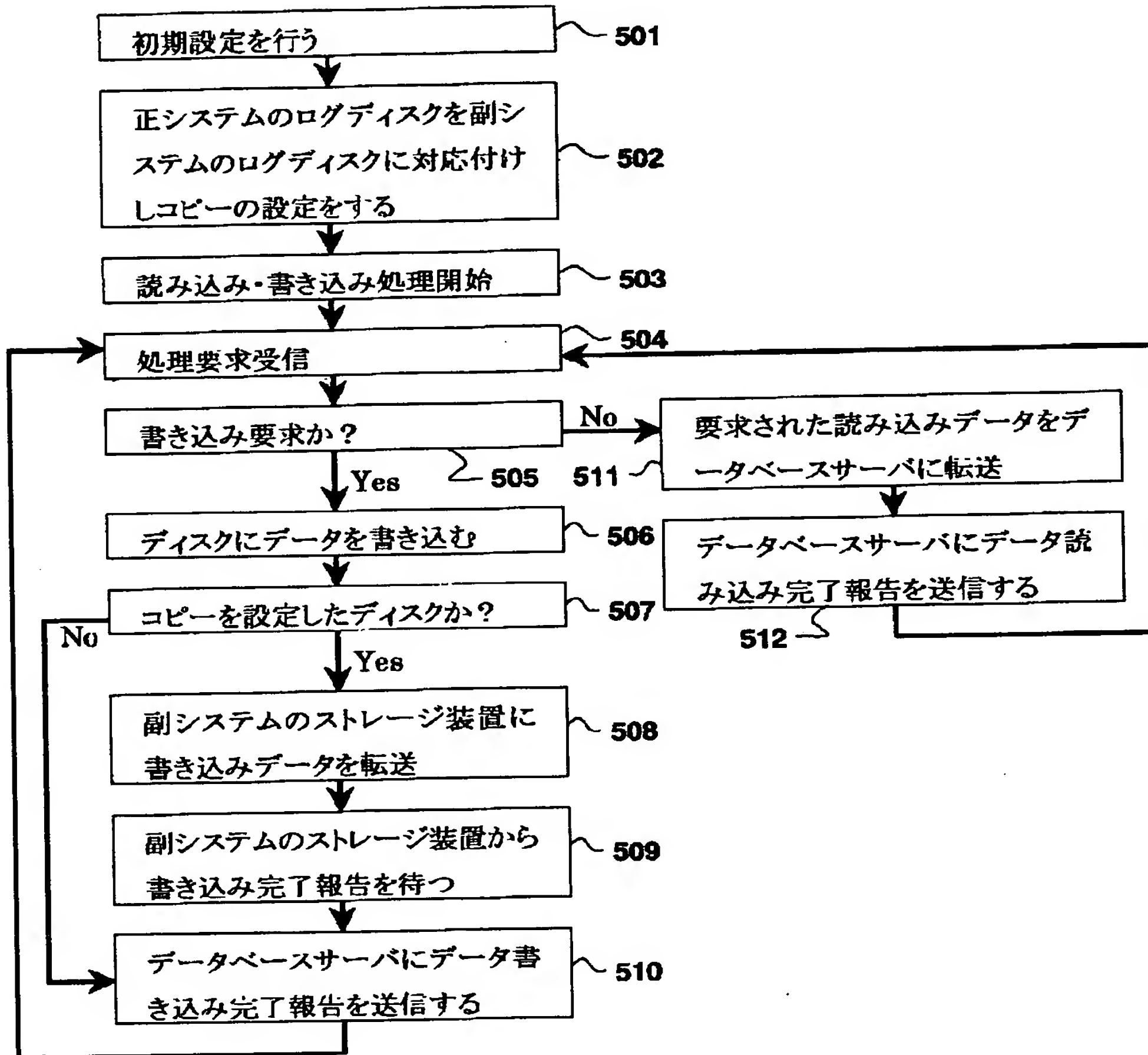
副システムデータベースサーバ処理手順



【図 5】

図5

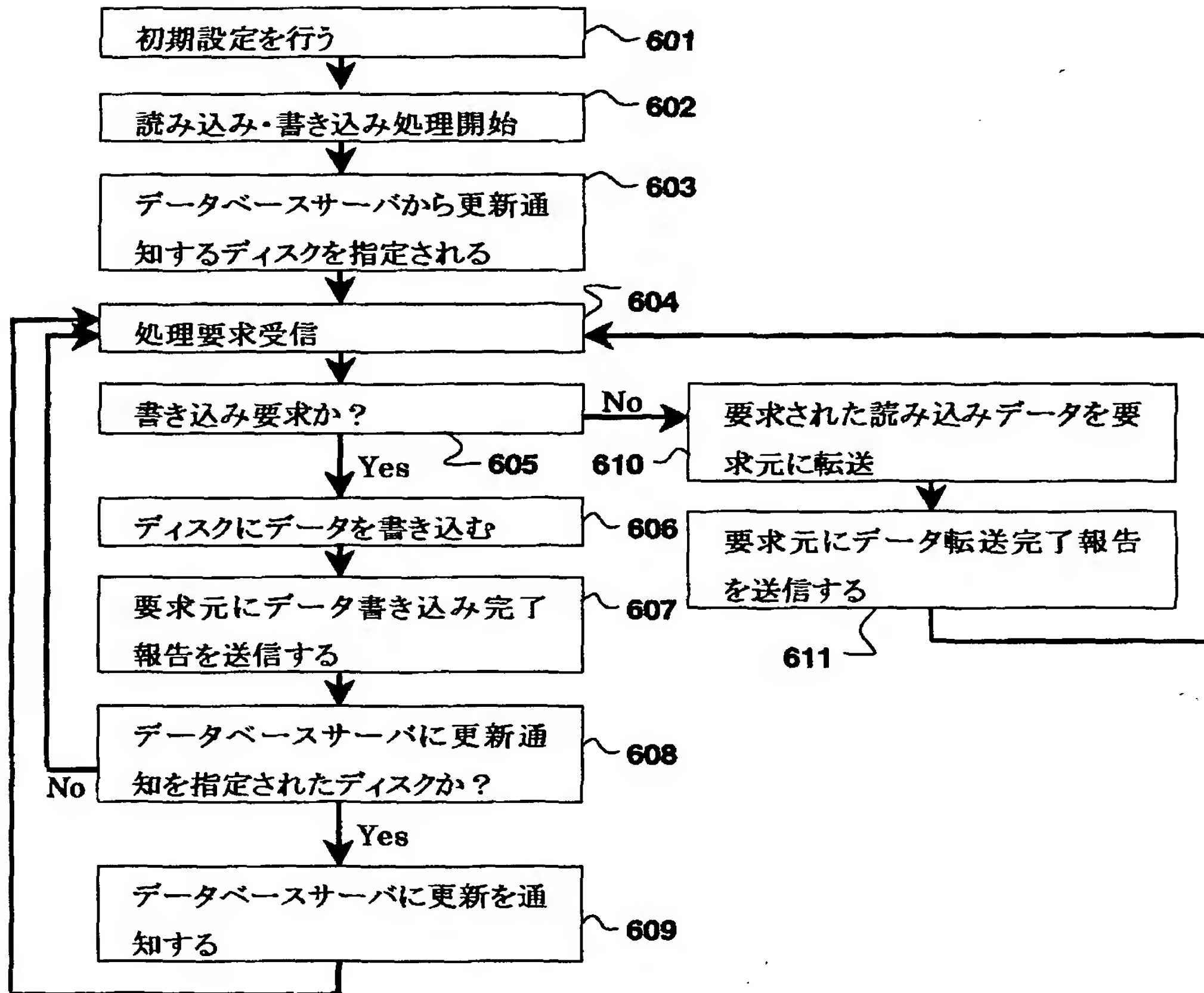
正システムストレージ装置処理手順



【図 6】

図6

副システムストレージ装置処理手順



【書類名】 要約書

【要約】

【課題】 データベースサーバおよびストレージ装置からなる業務処理システムにおいて、障害・災害・保守等により停止せざるを得ない場合に備え、データベースのデータ複製を行うために、通常業務を実行するデータベースサーバおよびストレージ装置の負荷を軽減し、また業務処理システムを予備システムに切り替える際に迅速に業務を引き継ぐことを可能とする。

【解決手段】 ストレージ装置により、DBMSのログディスクの複製を行い、切り替えを行うための予備システムのデータベースサーバにログディスクの更新をストレージ装置から通知することで、DBMSがログを参照してロールフォワードを行う。

【効果】 データの複製に必要であった業務処理システムのデータベースサーバおよびストレージ装置の負荷を軽減する。さらに、予備システムへの切り替えを迅速に実行することが可能となる。

【選択図】 図 1

認 定 ・ 付 加 情 報

特許出願の番号	特願 2 0 0 3 - 0 8 6 9 2 0
受付番号	5 0 3 0 0 5 0 0 7 4 9
書類名	特許願
担当官	第七担当上席 0 0 9 6
作成日	平成 1 5 年 3 月 2 8 日

< 認定情報・付加情報 >

【提出日】 平成15年 3月27日

出 願 人 履 歴 情 報

識別番号 [ 0 0 0 0 5 1 0 8 ]

1. 変更年月日 1 9 9 0 年 8 月 3 1 日

[変更理由] 新規登録

住 所 東京都千代田区神田駿河台 4 丁目 6 番地

氏 名 株式会社日立製作所